An Intelligent Credit Assessment System by Kernel Locality Preserving Projections and Manifold-Regularized SVM Models

Shian-Chang Huang

***Abstract*—Support vector machines (SVM) have been successfully applied in numerous areas of pattern recognitions, and have demonstrated excellent performance. However, traditional SVM does not make efficient use of both labeled training data and unlabeled testing data. Moreover, one usually encounters high dimensional and nonlinear distributed data in classification problems, especially in financial credit rating assessments. They generally degrade the performance of a classifier due to the curse of dimensionality. This study addresses these problems by proposing a novel intelligent system which integrates a kernel locality preserving projection (KLPP) with a data-dependent manifold-regularized SVM. KLPP is employed to gain a perfect approximation of data manifold and simultaneously preserve local within-class geometric structures according to prior class-label information. Empirical results indicate that, compared with other dimensionality reduction methods and conventional classifiers, the hybrid classifier performs best.**

***Index Terms*—Credit rating, dimensionality reduction, kernel locality preserving projections, subspace analysis, semi-supervised SVM.**

I. Introduction

The subprime mortgage crisis in 2007 results mainly from credit risk. Credit quality assessment is important for the banking sector. The bank with the most accurate estimation of its borrower's credit quality will be the most profitable. On the other hand, corporate credit quality (or rating) is typically very costly to assess, since they require agencies such as Standard and Poors or Moody to invest heavily in terms of time and human resources to perform deep analysis of a company's risk status. For controlling credit risk, all banking and investment institutes invest heavily on establishing an automatic decision support system for evaluating the credit quality of their borrowers. The objective of this study is thus to develop a reliable and accurate data mining system for credit quality evaluations.

Corporate credit rating predictions are studied intensively by the academic and business community. Many researchers have attempted to construct automatic classification systems using methods from statistics, data mining, and artificial intelligence. However, traditional methods usually perform poor when they encounter the high dimensional and nonlinear distributed financial input data. This study addresses the problems by integrating a kernel locality preserving projections with a semi-supervised version of support vector machine for credit rating forecasting.

Numerous classification techniques have been adopted for credit scoring. These techniques include: (1) traditional statistical methods; such as discriminant analysis, logistic regressiones [1], [2], and Bayesian networks; (2) non-parametric statistical models such as k-nearest neighbors [3]; (3) decision trees [4]; (4) neural networks [5], [6]. Recently, the support vector machine (SVM) [7]-[9], a special form of kernel classifiers, has become increasingly popular. The formulation of SVM simultaneously embodies the structural risk (a maximum margin classifier) and empirical risk minimization principles. Consequently, SVM combines excellent generalization properties with a sparse model representation.

Traditional SVM uses only the labeled data to train the model. However, unlabeled (testing) data usually provide important information about intrinsic geometry of the data manifold which helps for the out-of-sample model generalization and preventing the problem of overfitting. As a result, an approach that is able to make better use of both labeled and unlabeled data for training and regularization [10] to improve recognition performance is of potentially great practical significance. The new solution to deal with the problem is the semi-supervised learning (or transductive learning) [11], which falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). In the last decades, semi-supervised learning has attracted an increasing amount of attention. Recently, there are considerable interest and succuss on semi-supervised learning algorithms [12]-[15]. Manifold-regularized SVM (MR-SVM) [12] is a novel framework for data-dependent geometric regularization which brings together ideas from the theory of regularization in reproducing kernel Hilbert spaces (RKHS), manifold learning and spectral methods. Specifically, the MR-SVM used data-dependent norm to warp the structure of the RKHS to reflect the underlying geometry of the data.

The power of kernel methods lies in the implicit use of a high dimensional RKHS induced by a positive semidefinite (PSD) kernel. Kernel classifiers map input data into a high dimensional RKHS where simple linear classification is performed. However, owing to the large amounts of data from public financial statements that can be used for corporate credit rating predictions, the large scale of input data makes kernel classifiers infeasible due to the curse of dimensionality ([16]). Consequently, first one needs to transform the input data space to a suitable low dimensional subspace that optimally represents the data. Regarding dimensionality reduction, linear algorithms such as principal

Manuscript received May 15, 2014; revised July 17, 2014. This work was supported by the Ministry of Science and Technology of Taiwan.

Shian-Chang Huang is with the Department of Business Administration, National Changhua University of Education, Changhua, Taiwan (e-mail: shhuang@cc.ncue.edu.tw).

DOI: 10.7763/IJMO.2014.V4.405

component analysis (PCA) and linear discriminant analysis (LDA) are the two most widely used methods due to their relative simplicity and effectiveness.

However, as indicated by [17], in many real world problems there is no evidence that the data is sampled from a linear subspace. This problem has motivated researchers to consider manifold-based techniques for dimensionality reduction. Recently, various manifold learning techniques such as ISOMAP [18], Locally Linear Embedding LLE, [19] and Laplacian Eigenmap [20], have been proposed to reduce the dimensionality of a fixed input data set while maximally preserving certain inter-point relationships. However, these methods are unsuitable for credit rating forecasting, because they cannot provide an explicit subspace mapping for a new test sample. To address this deficiency, [21] proposed locality preserving projections (LPP) to approximate the eigenfunctions of the Laplace Beltrami operator on the data manifold, and so that new test samples can be easily mapped to the learned low-dimensional feature subspace. Although LPP is often effective, it performs poorly when data samples are subject to complex nonlinear changes since it is a linear method in nature.

This research adopted a kernel version of LPP KLPP, [22], [23] for subspace learning, which preserves geometric relations according to prior class-label information and represents complex nonlinear variations of real data by nonlinear kernel mapping. Incorporating KLPP significantly reduces the computational loading of kernel classifiers and simultaneously enhances forecasting accuracy. Moreover, this study also applies four types of multi-class kernel classifiers to classify enterprize credit ratings for comparison. Empirical results indicate that, compared with other dimensionality reduction methods and conventional classifiers, the hybrid classifier (KLPP+semi-supervised SVM) performs best. The proposed method can help financial institutions to accurately assess credit risk and substantially reduce losses.

The remainder of this paper is organized as follows: section II describes traditional locality preserving projections and multi-class kernel classifiers. Section III introduces the KLPP algorithm and manifold-regularized SVMs. Subsequently, Section IV describes the study data and discusses the empirical findings. Conclusions are given in Section V.

II. Prior Research

A. Locality Preserving Projections

This section presents the dimensionality reduction method of [21]. Given m samples $\boldsymbol{x}_i|_{i=1}^m \in \boldsymbol{R}^n$, dimensionality reduction aims at finding $\boldsymbol{z}_i|_{i=1}^m \in \boldsymbol{R}^d$, $d = n$, where $\boldsymbol{z}_i$ can represents $\boldsymbol{x}_i$. Locality Preserving Projection (LPP, [21]) is one of the famous algorithms. It builds a graph incorporating neighborhood information of the data set. Using the notion of Laplacian of the graph [24], one then computes a transformation matrix which maps the data points to a subspace. This transformation optimally preserves local neighborhood information in a certain sense. Based on standard spectral graph theory (see [24] for a comprehensive reference and [20] for applications to data representation), given a graph G with m vertices, each vertex represents a data point. Let W be a symmetric $m \times m$ matrix with W_{ij} having the weight of the edge joining vertices i and j. The G and W can be defined to characterize certain statistical or geometric properties of the data set. The purpose of LPP is to represent each vertex of a graph as a low dimensional vector that preserves similarities between the vertex pairs, where similarity is measured by the edge weight. Let $\boldsymbol{z} = [z_1, z_2, ..., z_m]^T$ be the map from the graph to the real line. The optimal z tries to minimize under appropriate constraint. This objective function incurs a heavy penalty if neighboring vertices i and j are mapped far apart. Therefore, minimizing it is an attempt to ensure that if vertices i and j are close then z_i and z_j are close as well. With some simple algebraic formulations, we have

$$\sum_{i,j}(z_i - z_j)^2 W_{i,j} \tag{1}$$

$$\sum_{i,j}(z_i - z_j)^2 W_{i,j} = 2\mathbf{z}^T L\mathbf{z}, \tag{2}$$

where $L = D - W$ is the graph Laplacian ([24]) and D is a diagonal matrix whose entries are column (or row, since W is symmetric) sums of W, $D_{ii} = \sum_j W_{ji}$. Finally, the minimization problem reduces to find

$$z^* = \arg\min_{\mathbf{z}^T D\mathbf{z}=1} z^T Lz = \operatorname{argmin} \frac{z^T Lz}{z^T Dz}. \tag{3}$$

The constraint $\mathbf{z}^T D\mathbf{z} = 1$ removes an arbitrary scaling factor in the embedding. The optimal Z s can be obtained by solving the minimum eigenvalue eigen-problem: $L\mathbf{z} = \lambda D\mathbf{z}$. If we choose a linear function, i.e., $z_i = f(\boldsymbol{x}_i) = \boldsymbol{a}^T \boldsymbol{x}_i$. Eq. (3) can be rewritten as:

$$\boldsymbol{a}^* = \arg\min_{\mathbf{z}^T D\mathbf{z}=1} \mathbf{z}^T L\mathbf{z} = \arg\min_{\boldsymbol{a}^T XDX^T \boldsymbol{a}=1} \boldsymbol{a}^T XLX^T \boldsymbol{a} \tag{4}$$

where $X = [x_1, ..., x_m]$. The optimal $\boldsymbol{a}$'s are the eigenvectors corresponding to the minimum eigenvalue of eigen-problem: $XLX^T \boldsymbol{a} = \lambda XDX^T \boldsymbol{a}$.

B. Support Vector Machines

Based on the structured risk minimization (SRM) principle, SVMs seek to minimize an upper bound of the generalization error instead of the empirical error as in other neural networks. SVM classifiers construct a hyperplane to separate the two classes (labelled $y \in \{-1, 1\}$) so that the margin (the distance between the hyperplane and the nearest point) is maximal. The SVM classification function is formulated as follows:

$$y = \operatorname{sign}(\boldsymbol{w}^T \phi(\boldsymbol{x}) + b), \tag{5}$$

where $\phi(x)$ is called the feature, which is a nonlinear mapping from the input space $\boldsymbol{x}$ to the future space. The coefficients $\boldsymbol{w}$ and b are estimated by the following optimization problem:

$$\min_{\mathbf{w},b} R(\mathbf{w},\xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\xi_i, \quad (6)$$

with

$$y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+b_i) \geq 1-\xi_i, \quad i=1,...,l \quad (7)$$

$$\xi_i \geq 0, \quad i=1,...,l, \quad (8)$$

where C is a prescribed parameter, which evaluates the trade-off between the empirical risk and the smoothness of the model.

After taking the Lagrangian and conditions for optimality, the dual solution of this convex optimization problem can be formulated as follows:

$$\max_{\alpha} D(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i,\mathbf{x}_j), \quad (9)$$

with constraints,

$$0 \leq \alpha_i \leq C, \quad i=1,...,l$$

$$\sum_{i=1}^{l}\alpha_i y_i = 0,$$

where α are Lagrangian multipliers, which are also the solution to the dual problem, and $K(\mathbf{x}_i,\mathbf{x}_j)$ is the kernel function. b follows from the complementarity Karush-Kuhn-Tucker (KKT) conditions. The decision function is given by

$$f(\mathbf{x}) = sign\left(\sum_{i=1}^{l}\alpha_i y_i K(\mathbf{x},\mathbf{x}_i) + b\right). \quad (12)$$

The value of the kernel is equal to the inner product of two vectors $\mathbf{x}$ and $\mathbf{x}_i$ in the feature space, such that $K(\mathbf{x},\mathbf{x}_i) = \phi(\mathbf{x})\phi(\mathbf{x}_i)$. Any function that satisfying Mercer's condition ([7]) can be used as the Kernel function.

C. General Kernel Classifiers: Kernel Dependency Estimation (KDE)

In [25]-[26], multi-class classifications are considered as structured output prediction problem. Both the input and the output portions of the data are mapped into their feature spaces, denoted as F and L, respectively. As usual, these mappings can be implicitly defined via two kernel functions:

$$K(\mathbf{x},\mathbf{x}') = \phi(\mathbf{x})\cdot\phi(\mathbf{x}')$$

$$L(\mathbf{y},\mathbf{y}') = \phi_L(\mathbf{y})\cdot\phi_L(\mathbf{y}'), \quad (14)$$

where $\phi: X \to T$ and $\phi_l: Y \to T$ are the input and the output feature mapping, respectively. Then every function $f: X \to T$ in the original domain, can be mapped to a corresponding function $F: K \to L$ in the transformed space defined as $F(\phi(\mathbf{x})) = \phi_L(f(\mathbf{x}))$.

Like distances between examples in the input space, it is also possible to think of loss function as a distance measure in the output space. We can measure inner products in the output space using a kernel function, denoting this as $L(\mathbf{y},\mathbf{y}') = \phi_L(\mathbf{y})\cdot\phi_L(\mathbf{y}')$, where $\phi_L: Y \to L$. This map makes it possible to consider a large class of nonlinear loss function. By using ϕ_L, one embeds the output objects in the space $\mathbb{L}$ for general loss measures. The general loss function make it possible to consider multi-class patterns, structured objects such as strings, trees, graphs and so forth as outputs.

First, define some kernel functions for output spaces. In M-class pattern recognition, given $Y = \{1,...,k\}$, one often uses the distance or loss function $LL(\mathbf{y},\mathbf{y}') = 1-[\mathbf{y}=\mathbf{y}']$, where $[\mathbf{y}=\mathbf{y}']$ is 1 if $\mathbf{y}=\mathbf{y}'$ and 0 otherwise. To construct a corresponding inner product it is necessary to embed this distance into a Euclidean space, which can be done using the following kernel:

$$L(\mathbf{y},\mathbf{y}') = \frac{1}{2}[\mathbf{y}=\mathbf{y}'], \quad (15)$$

as

$$LL(\mathbf{y},\mathbf{y}') = \|\phi_L(\mathbf{y}) - \phi_L(\mathbf{y}')\|^2 = L(\mathbf{y},\mathbf{y}) + L(\mathbf{y}',\mathbf{y}') - 2L(\mathbf{y},\mathbf{y}') = 1-[\mathbf{y}=\mathbf{y}'] \quad (10)$$

Similarly, one can also embed the input space X to a feature space F by another kernel function $\phi: X \to F$ for general distance measures. Our objective is to minimize the following risk function using the feature space F induced by the kernel K and the loss function measured in the space L nduced by kernel L, (11)

$$R(\alpha) = \int_{x\times y} LL(\mathbf{y}, f(\mathbf{x},\alpha))dP(\mathbf{x},\mathbf{y}), \quad (16)$$

where P is the joint distribution of $\mathbf{x}$ and $\mathbf{y}$. To do this we must learn the mapping from $\phi(\mathbf{x})$ to $\phi_L(\mathbf{y})$. Our solution is to decompose $\phi_L(\mathbf{y})$ into p orthogonal directions using kernel principal components analysis (see, e.g. [10] chapter 14). One can then learn the mapping from $\phi(\mathbf{x})$ to each direction independently using a standard kernel regression method. Finally, to output a estimation $\mathbf{y}$ given a test example $\mathbf{x}$ one must solve a pre-image problem. For more details concerning the algorithm we refer to [26].

III. The Proposed Method

A. Kernel Locality Preserving Projections

Since Locality Preserving Projection (LPP) is a linear method in nature, it is inadequate for representing nonlinear data space. Moreover, LPP seeks to preserve local structures (13) defined by the nearest neighbors. It often fails to preserve within-class local structure, which is very important for rating classification, because the nearest neighbors may belong to different classes due to the influence of complex variations. This paper uses a kernel version of LPP proposed by [22], [23] for subspace learning. First, the nonlinear kernel mapping is used to map the data into an implicit feature (RKHS) space F, and a linear transformation is then performed to preserve within-class geometric structures in F. Thus, a nonlinear subspace is obtained to approximate the intrinsic geometric structure of the data manifold. Namely, a function is selected in the high dimensional RKHS, i.e.,

$$\begin{aligned} \boldsymbol{z}_i &= f(\boldsymbol{x}_i) = P_\phi^T \phi(\boldsymbol{x}_i) = \sum_{j=1}^{m} \alpha_j \phi(\boldsymbol{x}_j)^T \phi(\boldsymbol{x}_i) \\ &= \sum_{j=1}^{m} \alpha_j K(\boldsymbol{x}_j, \boldsymbol{x}_i), \end{aligned} \tag{17}$$

where ϕ is the nonlinear mapping function, $K(\boldsymbol{x}_j, \boldsymbol{x}_i)$ is the Mercer Kernel, and P_ϕ is a projecting transformation that can preserve the within-class geometric structure of the data. Equations (3) of previous section can be rewritten as

$$\alpha^* = \arg\min_{\boldsymbol{z}^T D\boldsymbol{z}=1} \boldsymbol{z}^T L\boldsymbol{z} = \arg\min_{\alpha^T KDK\alpha=1} \alpha^T KLK\alpha, \tag{18}$$

where $\alpha = [\alpha_1, ..., \alpha_m]^T$. The optimal α 's are the eigenvectors corresponding to the minimum eigenvalue of the following eigen-problem: $KLK\alpha = \lambda KDK\alpha$.

The weight matrix W in Equation (1) is still unspecified. In [21] and [27], the weight matrix W was simply defined by nearest-neighbor relations. Here, similar to [28] prior class-label information is used to define the W . In fact, each entry in the weight matrix W can be regarded as the similarity metric of a pair of samples. The dot product between two samples is in a sense a similarity measure. Therefore, weight matrix W is defined as follows:

$$W_{ij} = \begin{cases} \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) & \text{if } \boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ both belong} \\ & \text{to the same class;} \\ 0 & \text{otherwise,} \end{cases}$$

That is, the within-class geometric information is emphasized, and the similarity between two samples is set to zero if they belong to different classes. Thus, matrices W and K are unified into a consistent dot product form except that matrix W has a strong constraint.

B. Manifold-Regularized SVM

Standard statistical learning assumes that there is a probability distribution P on $X \times Y$ according to which training examples are generated. Labeled examples are (x, y) pairs drawn from P . Unlabeled examples are simply $x \in X$ drawn from the marginal distribution P_X of P . Knowledge of the marginal P_X can give a better function learning for classification. Semi-supervised learning (especially, manifold regularization) assumes that if two points $x_1, x_2 \in X$ are close in the intrinsic geometry of P_X , then the conditional distributions $P(y \mid x_1)$ and $P(y \mid x_2)$ are similar. In other words, the conditional probability distribution $P(y \mid x)$ varies smoothly along the geodesics in the intrinsic geometry of P_X .

Given a set of labeled examples $(x_i, y_i), i = 1, ..., l$, the kernel learning framework estimates an unknown function in RKHS H_K with corresponding norm $\|\cdot\|_K$ by minimizing

$$f^* = \arg\min_{f \in H_K} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma \| f \|_K^2, \tag{19}$$

where V is some loss function, such as soft margin loss function for SVM. Penalizing the RKHS norm imposes smoothness conditions on possible solutions. The classical representer Theorem states that the solution to this minimization problem exists in H_K (with Mercer kernel K) and can be written as

$$f^* = \sum_{i=1}^{l} \alpha_i K(x_i, x), \tag{20}$$

The problem is reduced to optimizing over the finite dimensional space of coefficients α_i . Manifold-based regularization extends this framework by incorporating additional information about the geometric structure of the marginal P_X by introducing an additional regularizer

$$f^* = \arg\min_{f \in \mathrm{H_K}} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma \| f \|_K^2 + \gamma_I \| f \|_I^2, \tag{21}$$

where $\| f \|_I^2$ is an appropriate penalty term that reflect the intrinsic structure of P_X . In the setting, γ controls the complexity of the function in the ambient space while γ_I controls the complexity of the function in the intrinsic geometry of P_X .

However, in most applications we do not know P_X . Therefore one needs to get empirical estimates of $\| f \|_I$. In order to get good empirical estimates, unlabeled examples should be added in the training process. When the support of P_X is a compact submanifold $M \subset X$. A natural choice for $\| f \|_I$ is $\int_\mathrm{M} \langle \nabla_\mathrm{M} f, \nabla_\mathrm{M} f \rangle$.

The term $\int_\mathrm{M} \langle \nabla_\mathrm{M} f, \nabla_\mathrm{M} f \rangle$ may be approximated on the basis of labeled and unlabeled data using the graph Laplacian. Given a set of labeled examples $\{(x_i, y_i)\}_{i=1}^{l}$ and a set of u unlabeled examples $\{x_j\}_{j=l+1}^{l+u}$, the term $\int_\mathrm{M} \langle \nabla_\mathrm{M} f, \nabla_\mathrm{M} f \rangle \approx \frac{1}{(u+l)^2} \tilde{f}^T L \tilde{f}$, where $\tilde{f} = [f(x_1), ..., f(x_{l+u})]^T$ and $L = D - W$ is the graph Laplacian ([24]). W_{ij} are the edge weights in the data adjacency graph. D is a diagonal matrix whose entries are column (or row, since W is symmetric) sums of W , $D_{ii} = \sum_{j=1}^{l+u} W_{ij}$. Finally, the term $\frac{1}{(u+l)^2}$ is a normalizing coefficient.

Similarly, by representer theorem ([12]), the solution of equation (21) has an expansion in terms of both labeled and unlabeled examples. $f^*(x) = \sum_{i=1}^{l+u} \alpha_i K(x_i, x)$. For the setting of SVM, the loss function is defined as $V(x_i, y_i, f) = \max(0, 1 - y_i f(x_i))$. Introducing slack

variables, using standard Lagrange Multiplier techniques, a similar solution like traditional SVM can be derived easily. For more details concerning the solution, we refer to [12] and [13].

TABLE I: WILCOXON TESTS ON THE DIFFERENCES AMONG KLPP, ICA, AND PCA (P-VALUE)

	MR-SVM +KLPP	1-vs-rest +ICA	1-vs-rest +PCA	1-vs-rest	MSVM	KDE
MR-SVM+KLPP	1.0000	0.0079	0.0079	0.0079	0.0079	0.0079
1-vs-rest+ICA	0.0079	1.0000	0.4206	0.1508	0.6905	0.8413
1-vs-rest+PCA	0.0079	0.4206	1.0000	0.6905	0.4206	0.2222
1-vs-rest	0.0079	0.1508	0.6905	1.0000	0.1111	0.1508
MSVM	0.0079	0.6905	0.4206	0.1111	1.0000	0.8413
KDE	0.0079	0.8413	0.2222	0.1508	0.8413	1.0000

TABLE II: WILCOXON TESTS ON THE DIFFERENCES AMONG KLPP AND RFE (P-VALUE)

	MR-SVM +KLPP	1-vs-rest +RFE 5	1-vs-rest +RFE 10	1-vs-rest +RFE 15	1-vs-rest +RFE 20
MR-SVM+KLPP	1.0000	0.0079	0.0079	0.0079	0.0079
1-vs-rest+RFE 5	0.0079	1.0000	0.6905	0.8413	0.4603
1-vs-rest+RFE 10	0.0079	0.6905	1.0000	0.4206	0.5476
1-vs-rest+RFE 15	0.0079	0.8413	0.4206	1.0000	0.0952
1-vs-rest+RFE 20	0.0079	0.4603	0.5476	0.0952	1.0000

IV. EXPERIMENTAL RESULTS AND ANALYSIS

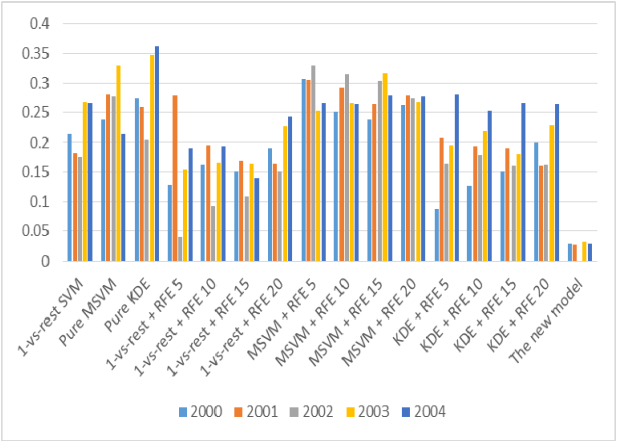
The TEJ (Taiwan Economic Journal) is a major provider of market data for Taiwan securities. This study used all the financial variables from the TEJ to forecast enterprize credit rating. Specifically, these financial variables include the following categories of information: company scale, financial structure, solvency, business performance, profitability, financial coverage and cash flow, for a total of thirty-six variables. Most of these variables are derived from publicly disclosed information that companies are required to file with authorities such as the securities and futures commission. They are important for financial analysis. Besides the financial variables, this study also included the historical rating of each company to improve rating accuracy.

Rating information for target companies was also obtained from the TEJ, which provides the credit rating for all publicly traded companies in Taiwan. A TEJ rating classifies the risk of a company not meeting its financial commitments over a one-year period as low, medium, or high. A low risk rating indicates that an organization has an extremely strong capacity to meet its commitments whereas a high risk rating indicates that an organization is likely to default.

This study tested three conventional classifiers and three kernel classifiers for corporate credit rating, including nearest neighbors (with one and three neighbors), logistic regressions, Bayesian networks, one-vs-rest SVM, multi-class SVM ([29]), and a general kernel classifier, kernel dependency estimation (KDE). For kernel classifiers, this study selected the polynomial kernel of two degrees for input owing to its good performance compared with other types of kernels. For output, a linear kernel is used for KDE due to its simplicity in computation. This study collected eighty-eight high technology companies that are traded on the Taiwan security market. Five ratings were obtained from TEJ for each company during the period from 2000 to 2004. The data set was randomly divided into ten parts, and ten-folds cross validation was applied to evaluate the model performance.

Fig. 1 shows the average error rates of all pure methods. On average, the one-vs-rest SVM is better than other classifiers. Consequently, this study implemented a manifold-regularized one-vs-rest SVM (MR-SVM) for subsequent classifications. The semi-supervised SVM makes efficient use of both labeled training data and unlabeled testing data to enhance the out-of-sample model generalization and to prevent the problem of overfitting.

TABLE III: PERFORMANCE COMPARISON (ERROR RATE) OF KLPP AND RFE FOR YEARS OF 2000-2004



On the other hand, due to the curse of dimensionality, irrelevant variables can degrade the performance of a kernel classifier. Suitable feature selection or dimensionality reduction schemes are often employed to improve classifier performance. Thus the KLPP algorithm were integrated into the MR-SVM, and compared with other two famous subspace learning algorithms, the PCA (Principal Component Analysis) and ICA (Independent Component Analysis, [30]). The dimension of subspace was set to five for all algorithms. Fig. 2 shows the results. Additionally, this study also compared KLPP with a famous feature selection algorithm, the recursive feature elimination (RFE) method proposed by [31]. The RFE algorithm recursively eliminates input variables to identify the most important five (RFE 5), ten (RFE 10), fifteen (RFE 15), and twenty (RFE 20) feature subsets for comparison. Fig. 3 shows the results. Fig. 2 and

Fig. 3 also show the performance of pure classifiers without any subspace learning and feature selection schemes for comparison.

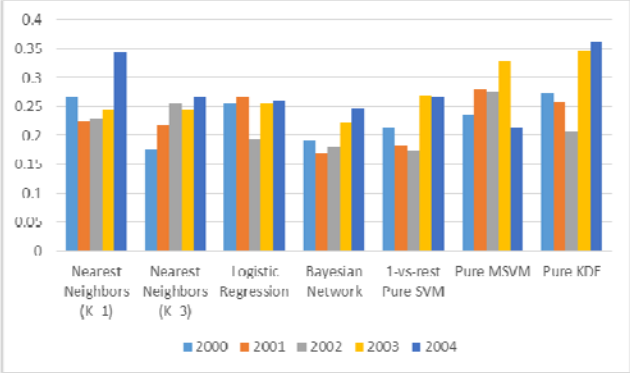


Fig. 1. Comparison of model forecasting performance (error rate) for years of 2000-2004.

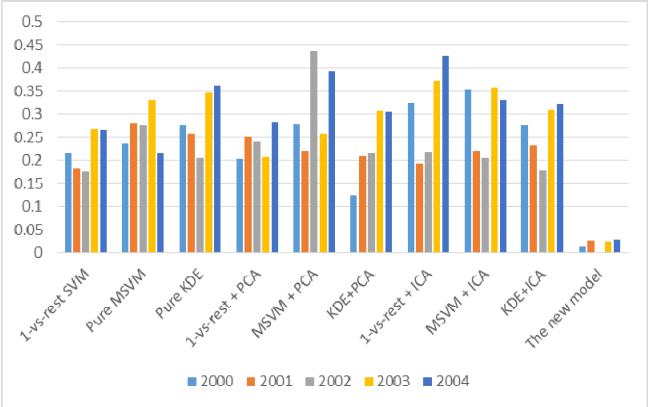


Fig. 2. Performance comparison (error rate) of three dimensionality reduction schemes for years of 2000-2004.

Fig. 2 shows that KLPP+MR-SVM significantly outperform other kernel classifiers. The MR-SVM with KLPP achieved the highest accuracy. This results fully demonstrate that in real rating problems the data is not sampled from a linear subspace. Hence, linear algorithms such as PCA and ICA fail to extract key information containing in the data. Considering graph-based nonlinear subspace learning (KLPP) and manifold-based semi-supervised SVM in rating problems are more effective.

The comparisons of KLPP with RFE in Fig. 3 indicate that MR-SVM with KLPP is the most cost-efficient model because it has the fewest dimensionality of subspace and achieves the best accuracy. Clearly, pure one-vs-rest SVM, MSVM, and KDE classifiers containing all of the input variables are less accurate than classifiers containing fewer variables. That is, more information does not necessarily improve accuracy.

Fig. 3 also shows that the performance improvement owing to RFE is limited regardless of the number of key features selected by RFE. RFE forms feature subset in original input space, but KLPP nonlinearly forms a subspace preserving local geometry of the within-class samples in high dimensional feature space, which contains sufficient information or latent structures to discriminate or represent the data, while the subset formed by RFE does not.

The Wilcoxon rank-sum test ([32]) is a nonparametric alternative (for sample median) to the two sample t-test which is based solely on the order in which the observations from the two samples fall. For performance comparison, the Wilcoxon rank-sum test is performed on different models. The Wilcoxon comparison of our new model with ICA and PCA based classifiers is displayed in Table I, while the Wilcoxon comparison of our new model with RFE based classifiers is displayed in Table II.

Tables I and II clearly demonstrate the superiority of the new hybrid classifier. Under 1% of significance, the new classifier substantially outperforms the ICA and PCA based model, moreover, it also outperforms RFE based model even higher dimension of features are selected.

V. Conclusions

Corporate credit ratings provide important information about credit risk for banks or investors in financial markets. This study integrated KLPP with MR-SVM to create a novel system for rating predictions. The performance of the new system was examined using a data set comprising a large amount of financial information regarding Taiwanese high technology companies. The empirical results showed that the proposed system is more accurate and robust than pure SVM classifiers, and also outperforms conventional techniques when applied to multiple-class credit rating problems.

Using the class information of data to guide the manifold learning, KLPP is a nonlinear subspace learning method, which preserves geometric relations according to prior class-label information and represents complex nonlinear variations of data samples by nonlinear kernel mapping. Integrating KLPP in a classifier can reduce its computational loading and simultaneously enhance their performance. In the second stage, this study used a manifold-regularized semi-supervised SVM for classification. The MR-SVM uses the data-dependent norm on RKHS to warp the structure of the RKHS to reflect the underlying geometry of the data. The success of the our hybrid classifier mainly attributes to the combination of two techniques.

Future research may consider input of other data such as non-financial and macroeconomic variables. However, including more information does not guarantee higher accuracy. In this situation, nonlinear subspace analysis is an important strategies for enhancing classifier performance. What types of supervised or semi-supervised subspace learning algorithm are more effective and efficient to incorporate into kernel classifiers need further study.

References

[1] A. Steenackers and M. J. Goovaerts, "A credit scoring model for personal loans," *Insurance Mathematics Economics*, vol. 8, pp. 31-34, 1989.

[2] M. Stepanova and L. C. Thomas, "PHAB scores: proportional hazards analysis behavioural scores," *The Journal of the Operational Research Society*, vol. 52, pp. 1007-1016, 2001.

[3] W. E. Henley and D. J. Hand, "Construction of a k-nearest neighbour credit-scoring system," *IMA Journal of Management Mathematics*, vol. 8, pp. 305-321, 1997.

[4] M. B. Yobas, J. N. Crook, and P. Ross, "Credit scoring using neural and evolutionary techniques," *IMA Journal of Management Mathematics*, vol. 11, pp. 111-125, 2000.

[5] V. S. Desai, J. N. Crook, and J. G. A. Overstreet, "A comparison of neural networks and linear scoring models in the credit union

environment," *European Journal of Operations Management*, vol. 95, pp. 24-37, 1996.
[6] D. West, "Neural network credit scoring models," *Computers and Operations Research*, vol. 27, pp. 1131-1152, 2000.
[7] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Second Edition, New York, Springer, 1999.
[8] N. Cristianini and J. S. Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.
[9] B. Schoelkopf, C. J. C. Burges, and A. J. Smola, *Advances in kernel methods - support vector learning*, MIT Press, Cambridge, MA, 1999.
[10] B. Scholkopf and A. J. Smola, *Learning with Kernels Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, M.A, 2002.
[11] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning,* MIT Press, Cambridge, M.A, 2006.
[12] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399-2434, 2006.
[13] V. Sindhwani, P. Niyogi, and M. Belkin, "Beyond the point cloud: from transductive to semi-supervised learning," in *Proc. 2005 Int. Conf. Machine Learning*, pp. 824-831, 2005.
[14] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," *Advances in Neural Information Processing Systems*, vol. 16, 2003.
[15] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proc. the twentieth Internation Conference on Machine Learning*, pp. 1-8, 2003.
[16] R. Bellman, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, 1961.
[17] S. Yan, D. Xu, B. Zhang, H. J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extension: A general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40-51, 2007.
[18] J. Tenenbaum, V. D. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319-2323, 2000.
[19] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323-2326, 2000.
[20] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Advances in Neural Information Processing Systems*, vol. 14, pp. 585-591, MIT Press, Cambridge, MA.
[21] X. He and P. Niyogi, "Locality preserving projections," A*dvances in Neural Information Processing Systems*, vol. 16, Vancouver, Canada, 2003.
[22] D. Cai, and X. He, and J. Han, "Spectral regression for efficient regularized subspace learning," in *Proc. IEEE International Conference on Computer Vision*, pp. 1-8, 2007.
[23] D. Cai, X. He, and J. Han, "Spectral Regression for Dimensionality Reduction," Department of Computer Science Technical Report No. 2856, University of Illinois at Urbana-Champaign, 2007.
[24] F. R. K. Chung, "Spectral graph theory," *Regional Conference Series in Mathematics*, vol. 92, 1997.
[25] C. Cortes, M. Mohri, and J. Weston, "A general regression technique for learning transductions," in *Proc. the 22nd international conference on Machine learning*, pp. 153-160, 2005.
[26] J. Weston, O. Chapelle, A. Elisseeff, B. Schölkopf, and V. Vapnik, "Kernel dependency estimation," *NIPS*, pp. 873-880, 2002.
[27] X. He, S. Yan, Y. Hu, and H. J. Zhang, "Learning a locality preserving subspace for visual recognition," in *Proc. Ninth International Conference on Computer Vision, France*, October 2003, pp. 385-392.
[28] J. Cheng, Q. Liu, H. Lu, and Y.W. Chen, "Supervised kernel locality preserving projections for face recognition," *Neurocomputing*, vol. 67, pp. 443-449, 2005.
[29] J. Weston and C. Watkins, *Support Vector Machines for Multi-Class Pattern Recognition*, ESANN'99, 1999.
[30] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley Interscience, 2001.
[31] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389-422, 2002.
[32] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics*, vol. 1, pp. 80-83, 1945.

Shian-Chang Huang received his MS degree in electric engineering from National Tsing Hwa University, and his PhD degree in financial engineering from National Taiwan University. He is currently a professor at the Department of Business Administration, National Changhua University of Education, Taiwan. His research interests include machine learning, soft computing, signal processing, data mining, computational intelligence, and financial engineering.